Attribute-Efficient Learning of Monomials over Highly-Correlated Variables

Alexandr Andoni, Rishabh Dudeja, Daniel Hsu, **Kiran Vodrahalli**

Columbia University

Yahoo Research, Aug. 2019

General Learning Problem

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f is from a low-complexity class

Assumption 2: $\boldsymbol{x}^{(i)} \sim D$, some reasonable distribution

Goal: Recover f exactly

A Natural Class?

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f depends on only k features

A Natural Class?

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f depends on only k features

Goal 1: Learn f with low sample complexity

Goal 2: Learn f computationally efficiently

A Natural Class?

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f depends on only k features

f linear → classical compressed sensing

Goal 2: Learn f computationally efficiently

Linear functions: Compressed Sensing

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f depends on only k features

Goal 1: Learn f with $poly(\log p, k)$ samples

Goal 2: Learn f in $poly(p, k, m)$ runtime

A Natural Class?

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

Assumption 1: f depends on only k features

f nonlinear ? Perhaps a polynomial…

Goal 2: Learn f computationally efficiently

Sparse polynomial functions

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \subset \mathbb{R}^{p} \times \mathbb{R}$, drawn i.i.d.

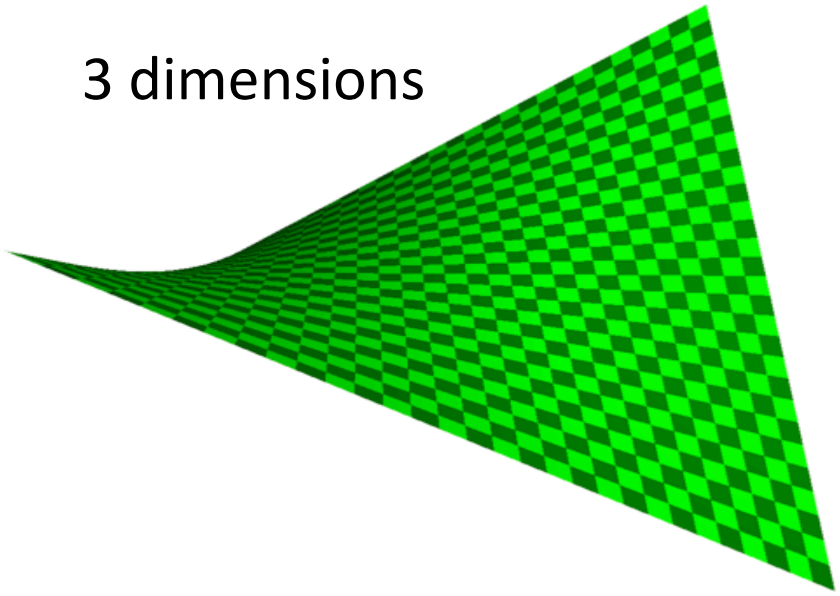
Assumption 1: f depends on only k features

Goal 1: Learn f with $poly(\log p, k)$ samples

Goal 2: Learn f in $poly(p, k, m)$ runtime

Simplest Case: Sparse Monomials

A Simple
Nonlinear
Function Class



In p dimensions
and k sparse

Ex: $f(x_1, \ldots, x_p) := \underbrace{x_3 \cdot x_{17} \cdot x_{44} \cdot x_{79}}_{k=4}$

The Learning Problem

Given: $\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m}$, drawn i.i.d.

Assumption 1: f is a k-sparse monomial function

Assumption 2: $\boldsymbol{x}^{(i)} \sim \mathcal{N}(0, \Sigma)$

Goal: Recover f exactly

Attribute-Efficient Learning

- Sample efficiency: $m = \mathrm{poly}(\log(p), k)$
- Runtime efficiency: $\mathrm{poly}(p, k, m)$ ops
- Goal: achieve both!

Motivation

$$x_i \in \{\pm 1\}$$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]

$$x_i \in \mathbb{R}$$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$x_i \in \{\pm 1\}$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$x_i \in \{\pm 1\}$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$$x_i \in \{\pm 1\}$$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime

$$x_i \in \mathbb{R}$$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$x_i \in \{\pm 1\}$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$x_i \in \{\pm 1\}$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Motivation

$$x_i \in \{\pm 1\}$$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$$x_i \in \mathbb{R}$$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]
  - $poly(p, 2^d, s)$ samples and runtime
    - d is maximum degree
    - s is number of monomials

Motivation

$$x_i \in \{\pm 1\}$$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum’98, Klivans&Servedio’06, Kalai+’09, Kocaoglu+’14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$$x_i \in \mathbb{R}$$

- Sparse **linear** regression
  [Candes+’04, Donoho+’04, Bickel+’09…]
- Sparse sums of monomials
  [Andoni+’14]
  - $poly(p, 2^d, s)$ samples and runtime
    - d is maximum degree
    - s is number of monomials
  - Works for Gaussian and Uniform distributions…

Motivation

$$x_i \in \{\pm 1\}$$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$$x_i \in \mathbb{R}$$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]
  - $poly(p, 2^d, s)$ samples and runtime
    - d is maximum degree
    - s is number of monomials
  - Works for Gaussian and Uniform distributions…
    - BUT they must be **product distributions**!

Motivation

$x_i \in \{\pm 1\}$

- Monomials ≡ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]
  - $poly(p, 2^d, s)$ samples and runtime
    - d is maximum degree
    - s is number of monomials
  - Works for Gaussian and Uniform distributions…
    - BUT they must be **product distributions**!
    - Whitening blows up complexity

Motivation

$x_i \in \{\pm 1\}$

- Monomials $\equiv$ Parity functions
- No attribute-efficient algs!
  [Blum'98, Klivans&Servedio'06, Kalai+'09, Kocaoglu+'14…]
  - Even in the **noiseless** setting
  - Brute force: $poly(\log p, k)$ samples, $O(p^k)$ runtime
    - Can improve to $O(p^{k/2})$ runtime
  - Improper learner: $O(p^{1-1/k})$ samples, $O(p^4)$ runtime
  - Even under avg. case assumptions…
    - $poly(p, 2^k)$ samples and runtime

$x_i \in \mathbb{R}$

- Sparse **linear** regression
  [Candes+'04, Donoho+'04, Bickel+'09…]
- Sparse sums of monomials
  [Andoni+'14]

Uncorrelated features:

$$\mathbb{E}[xx^T] = \begin{pmatrix} \sigma_1^2 & & & & & 0 \\ & \sigma_2^2 & & & & \\ & & \sigma_3^2 & & & \\ & & & \sigma_4^2 & & \\ & & & & \sigma_5^2 & \\ 0 & & & & & \sigma_6^2 \end{pmatrix}$$

Motivation

$x_i \in \{\pm 1\}$

- Monomials
- No attribut
  [Helmbold+ '92, B
  Kalai+'09, Kocaog

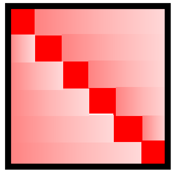
$x_i \in \mathbb{R}$

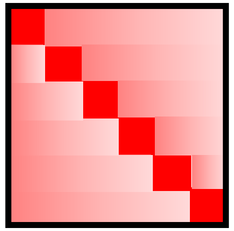
- regression
  +'04, Bickel+'09…]
- omials
- ated features:

$$\begin{pmatrix} \sigma_1^2 & & & & & \\ & \sigma_2^2 & & & 0 & \\ & & \sigma_3^2 & & & \\ & & & \sigma_4^2 & & \\ & 0 & & & \sigma_5^2 & \\ & & & & & \sigma_6^2 \end{pmatrix}$$

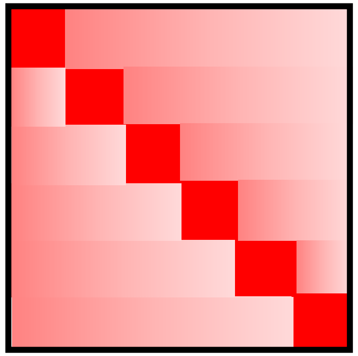
Question: What if

$$\mathbb{E}[xx^T] = \begin{pmatrix} 1 & & & & & \\ & 1 & & \leq \rho & & \\ & & 1 & & & \\ & & & 1 & & \\ & \leq \rho & & & 1 & \\ & & & & & 1 \end{pmatrix} \;?$$

Potential Degeneracy of  $= \mathbb{E}[xx^T]$

Ex: $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_p \end{bmatrix} \sim \begin{bmatrix} \mathcal{N}(0,1) \\ \mathcal{N}(0,1) \\ (x_1 + x_2)/\sqrt{2} \\ \mathcal{N}(0,1) \\ \vdots \\ \mathcal{N}(0,1) \end{bmatrix}$ $\longrightarrow$  $= \begin{bmatrix} 1 & 0 & \sqrt{.5} & & & \\ 0 & 1 & \sqrt{.5} & & \mathbf{0} & \\ \sqrt{.5} & \sqrt{.5} & 1 & & & \\ & & & \ddots & & \\ & \mathbf{0} & & & 1 & 0 \\ & & & & 0 & 1 \end{bmatrix}$

Singular matrix

 can be **low-rank!**

Rest of the Talk

1. Algorithm

2. Intuition

3. Analysis

4. Conclusion

1. Algorithm

The Algorithm

Ex: $f(x_1, \dots, x_p) \coloneqq x_3 \cdot x_{17} \cdot x_{44} \cdot x_{79}$

Step 1

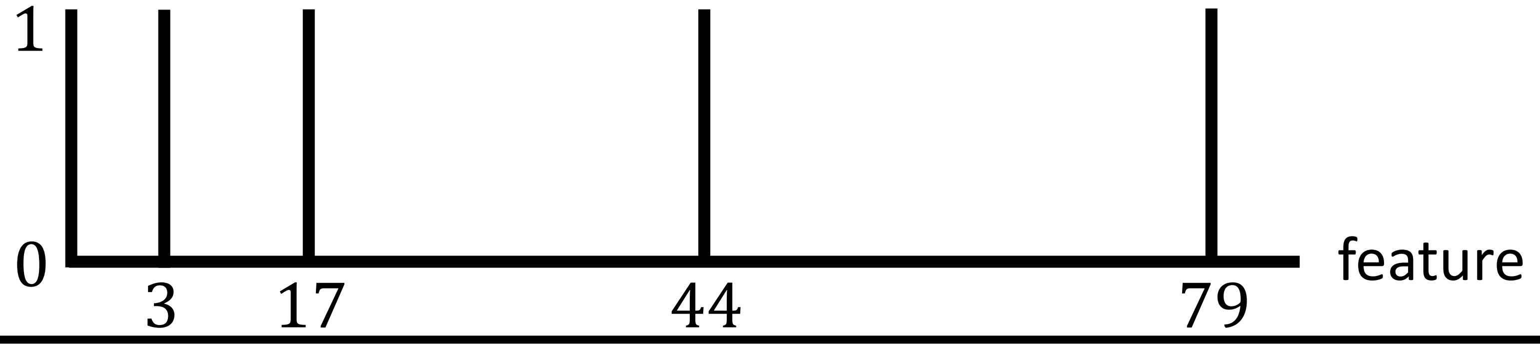
$$\left\{\left(\boldsymbol{x}^{(i)}, f\left(\boldsymbol{x}^{(i)}\right)\right)\right\}_{i=1}^{m} \xrightarrow{\log|\cdot|} \left\{\left(\log\left|\boldsymbol{x}^{(i)}\right|, \log\left|f\left(\boldsymbol{x}^{(i)}\right)\right|\right)\right\}_{i=1}^{m}$$

Gaussian Data → Log-transformed Data

Step 2

Sparse Regression:

(Ex: Basis Pursuit)



2. Intuition

Why is our Algorithm Attribute-Efficient?

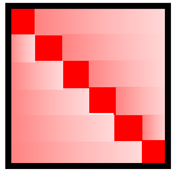
- Runtime: basis pursuit is efficient

- Sample complexity?
  - Sparse **linear** regression? E.g.,

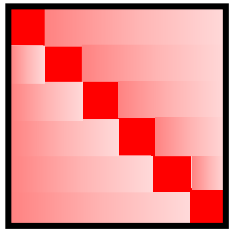
$$\log|f(x_1, \dots, x_p)| := \log|x_3| + \log|x_{17}| + \log|x_{44}| + \log|x_{79}|$$

  - But: sparse recovery properties may not hold…

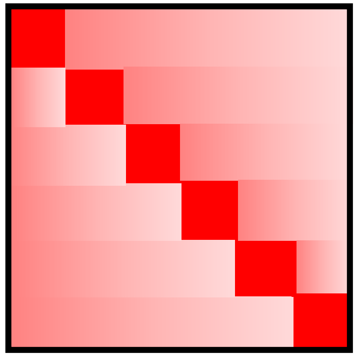
Degenerate High Correlation



$= \mathbb{E}[xx^T]$

Recall the example: 

$$= \begin{bmatrix} 1 & 0 & \sqrt{.5} & & & \\ 0 & 1 & \sqrt{.5} & & \mathbf{0} & \\ \sqrt{.5} & \sqrt{.5} & 1 & & & \\ & & & \ddots & & \\ & \mathbf{0} & & & 1 & 0 \\ & & & & 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} -1/2 \\ -1/2 \\ 1/\sqrt{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

3-sparse

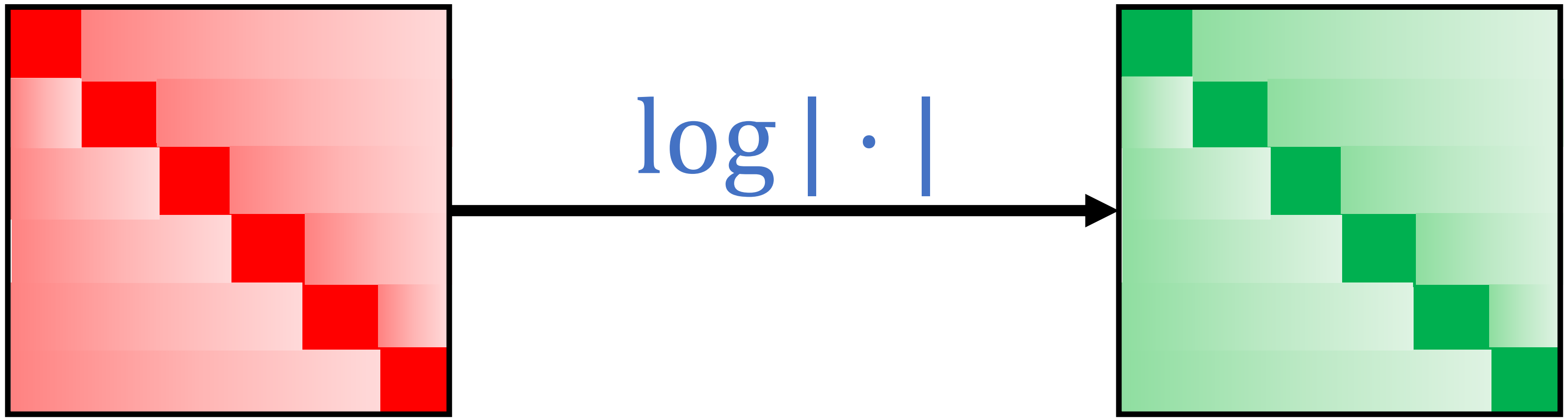
0-eigenvectors can be k-sparse

Sparse recovery conditions false!

Summary of Challenges

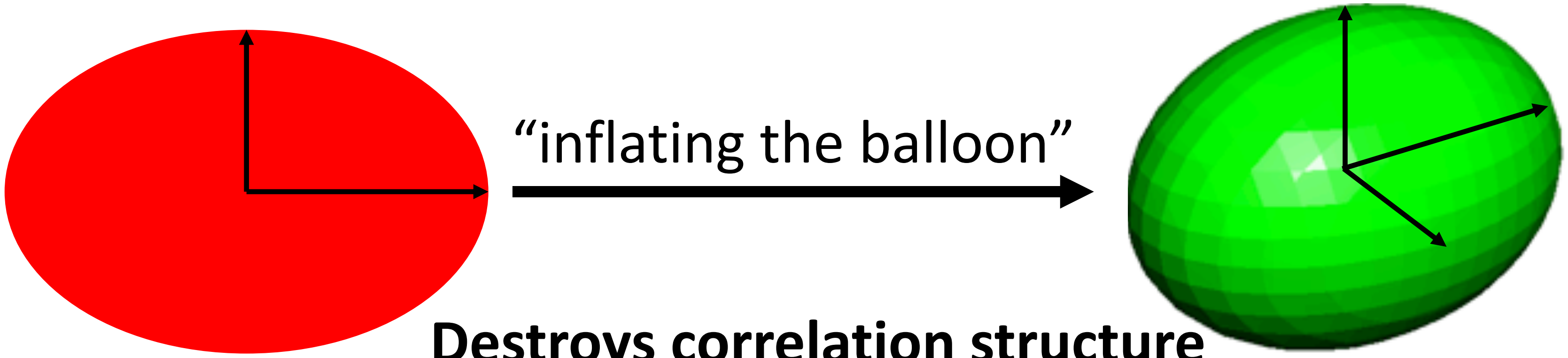
- Highly correlated features
- Nonlinearity of $\log|\cdot|$
- Need a recovery condition…

Log-Transform affects Data Covariance



$\mathbb{E}[xx^T] \succcurlyeq 0$ $\qquad$ $\mathbb{E}[\log|x| \log|x|^T] \succ 0$

Spectral View:



Destroys correlation structure

3. Analysis

Restricted Eigenvalue Condition [Bickel, Ritov, & Tsybakov '09]

Restricted Eigenvalue $RE(k)$

$$\min_{v \in C} \frac{v^T X X^T v}{||v||_2^2} > \epsilon$$

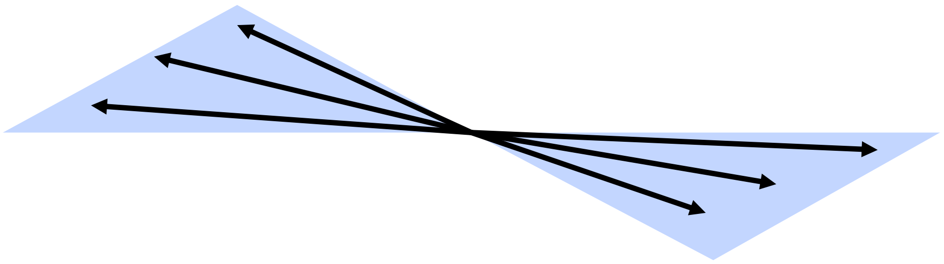
"restricted strong convexity"

Note: $RE(k) \geq \lambda_{min}(XX^T)$

Ex: $S = \{3, 17, 44, 79\}$

$k = 4$

Cone restriction



$C = \{v \colon ||v_S||_1 \geq ||v_{S^c}||_1\}$

$|S| = k$

Restricted Eigenvalue Condition [Bickel, Ritov, & Tsybakov '09]

Restricted Eigenvalue $RE(k)$

$$\min_{v \in C} \frac{v^T X X^T v}{||v||_2^2} > \epsilon$$

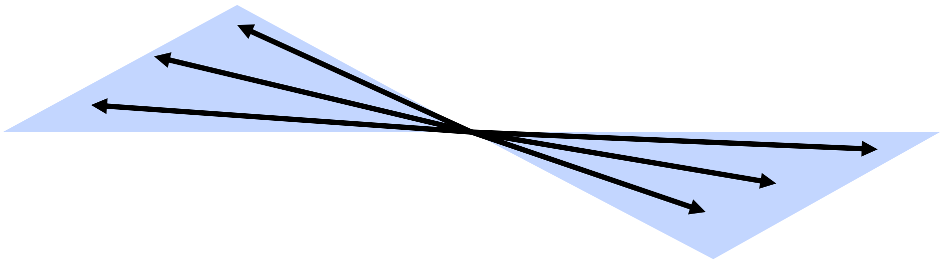
"restricted strong convexity"

Note: $RE(k) \geq \lambda_{min}(XX^T)$

Sufficient to prove exact recovery for basis pursuit!

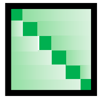
Ex: $S = \{3, 17, 44, 79\}$
$k = 4$

Cone restriction

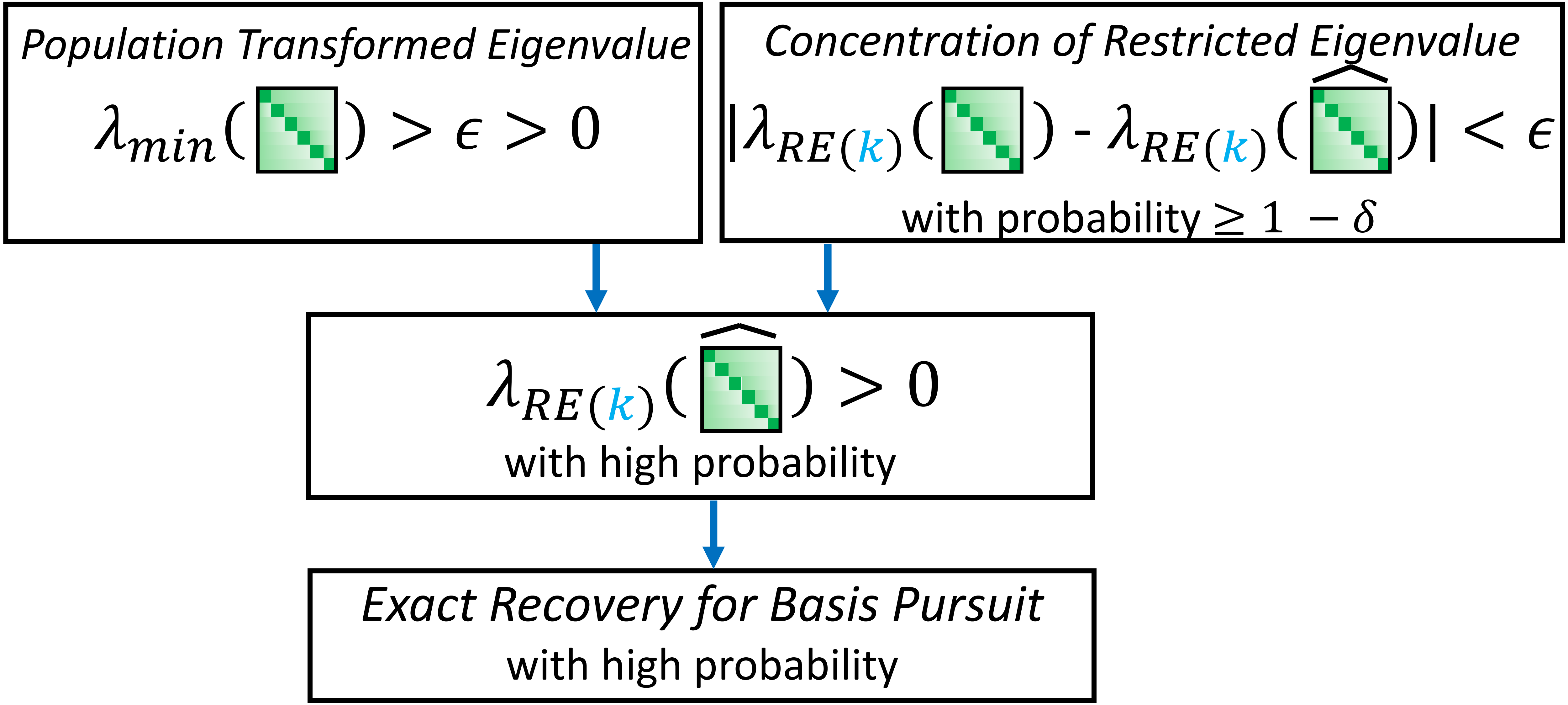


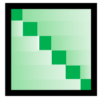
$C = \{v\colon ||v_S||_1 \geq ||v_{S^c}||_1\}$

$|S| = k$

 $= \mathbb{E}[\log|x| \log|x|^T]$

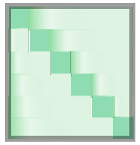
Sample Complexity Analysis



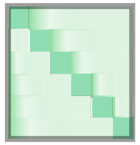
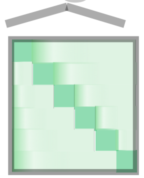
 $= \mathbb{E}[\log|x| \log|x|^T]$

Sample Complexity Analysis

Population Transformed Eigenvalue

$\lambda_{min}($  $) > \epsilon > 0$

Concentration of Restricted Eigenvalue

$|\lambda_{RE(k)}($  $) - \lambda_{RE(k)}($  $)| < \epsilon$

ility $\geq 1 - \delta$

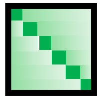
Sample Complexity Bound:

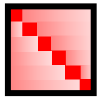
$$m = \tilde{O}\left(\frac{k^2 \log 2k}{1-\rho} \cdot \log^2 \frac{2p}{\delta}\right)$$

with high probability

Exact Recovery for Basis Pursuit

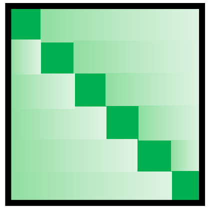
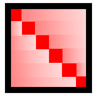
with high probability

 $= \mathbb{E}[\log|x| \log|x|^T]$

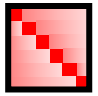
 $= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

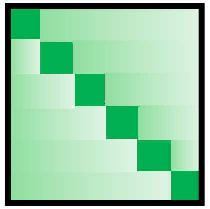
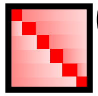
- Hermite expansion of $\log|\cdot|$:

 $= c_0^2 1_{pxp} + \sum_{l=1}^{\infty} c_{2l}^2$  $^{(2l)}$

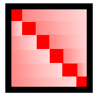
- $l \geq 1$: $c_{2l}^2 \sim \frac{\sqrt{\pi}}{4} \cdot \frac{1}{l^{3/2}}$

-  $^{(2l)}$ off-diagonals decay fast!

- Apply λ_{min} to Hermite formula:

λ_{min}  $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$  $^{(2l)}$

- Apply Gershgorin Circle Theorem:

λ_{min}  $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

(for large enough l)

Population Minimum Eigenvalue

$= \mathbb{E}[\log|x| \log|x|^T]$

$= \mathbb{E}[xx^T]$

- Hermite expansion of $\log|\cdot|$:

$$= c_0^2 1_{pxp} + \sum_{l=1}^{\infty} c_{2l}^2 \,{}^{(2l)}$$

- $l \geq 1$: $c_{2l}^2 \sim \frac{\sqrt{\pi}}{4} \cdot \frac{1}{l^{3/2}}$

- $^{(2l)}$ off-diagonals decay fast!

$$\mathbb{E}_a[(\log|a|)^2] < \infty$$

+

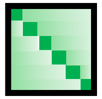
$$\log|a| = \sum_{l=0}^{\infty} c_l H_l(a)$$

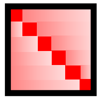
+

$$\mathbb{E}_{a,a'}[H_l(a) H_{l'}(a')] = \rho_{a,a'}^l$$

if $l = l'$, 0 otherwise

(for large enough l)

 $= \mathbb{E}[\log|x| \log|x|^T]$

 $= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

- Hermite expansion of $\log|\cdot|$:

 $= c_0^2 1_{pxp} + \sum_{l=1}^{\infty} c_{2l}^2$ $^{(2l)}$

- $l \geq 1$: $c_{2l}^2 \sim \frac{\sqrt{\pi}}{4} \cdot \frac{1}{l^{3/2}}$

- $^{(2l)}$ off-diagonals decay fast!

- Apply λ_{min} to Hermite formula:

Integration by Parts

+

Recursive Properties of Hermite Polynomials

+

Stirling Approximation

Note: $c_l = 0$ if l odd.

$= \mathbb{E}[\log|x| \log|x|^T]$

$= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

- Hermite expansion of $\log|\cdot|$:

$$= c_0^2 1_{pxp} + \sum_{l=1}^{\infty} c_{2l}^2 {}^{(2l)}$$

- $l \geq 1$: $c_{2l}^2 \sim \frac{\sqrt{\pi}}{4} \cdot \frac{1}{l^{3/2}}$

- ${}^{(2l)}$ off-diagonals decay fast!

- Apply λ_{min} to Hermite formula:

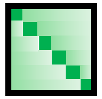
$$\lambda_{min} \geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min} {}^{(2l)}$$

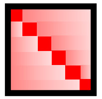
- Apply Gershgorin Circle Theorem:

$$\lambda_{min} {}^{(2l)} \geq 1 - (p-1)\rho^{2l}$$

Elementwise Matrix Product

Population Minimum Eigenvalue

 $= \mathbb{E}[\log|x| \log|x|^T]$

 $= \mathbb{E}[xx^T]$

- Hermite expansion of $\log|\cdot|$:

λ_{min} is superadditive

$+$

$$\lambda_{min}(1_{pxp}) = 0$$

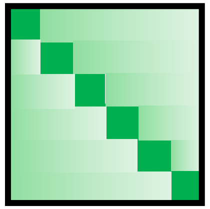
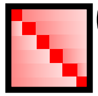
$+$

$$\lambda_{min}(A^{(l)}) \geq \lambda_{min}(A)$$

when A is symmetric PSD with 1 on the diagonal [Bapat & Sunder, '85]

- $^{(2l)}$ off-diagonals decay fast!

- Apply λ_{min} to Hermite formula:

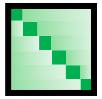
λ_{min}  $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$  $^{(2l)}$

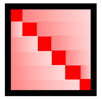
- Apply Gershgorin Circle Theorem:

λ_{min} $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

(for large enough l)

Population Minimum Eigenvalue

 $= \mathbb{E}[\log|x| \log|x|^T]$

 $= \mathbb{E}[xx^T]$

- Hermite expansion of $\log|\cdot|$:

 $= c_0^2 1_{pxp} + \sum_{l=1}^{\infty} c_{2l}^2$ $^{(2l)}$

- $l \geq 1$: $c_{2l}^2 \sim \frac{\sqrt{\pi}}{4} \cdot \frac{1}{l^{3/2}}$

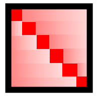
- $^{(2l)}$ c… …ast

Choose l so that $(p-1)\rho^{2l} < 1$

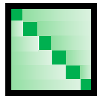
- Apply λ_{min} to Hermite formula:

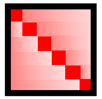
λ_{min} $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$ $^{(2l)}$

- Apply Gershgorin Circle Theorem:

λ_{min}  $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

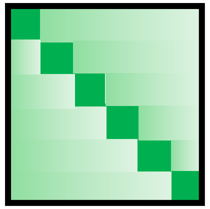
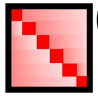
(for large enough l)

 $= \mathbb{E}[\log|x| \log|x|^T]$

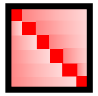
 $= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

- Apply λ_{min} to Hermite formula:

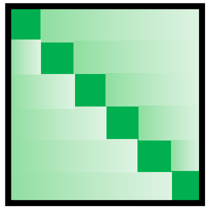
λ_{min}  $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$  $^{(2l)}$

- Apply Gershgorin Circle Theorem:

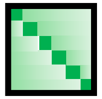
λ_{min}  $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

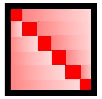
(for large enough l)

The Full, Ugly Bound

λ_{min}  $\geq$

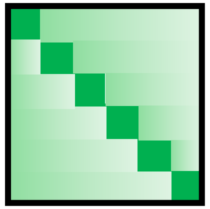
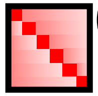
$\sum_{l=1}^{\frac{\log(p-1)}{2\log(\rho^{-1})}} \frac{\lambda_{min}$  $^{(2l)}}{5l^{3/2}} + \frac{2}{5}\sqrt{\frac{2\log\rho^{-1}}{\log\frac{p-1}{\rho} + \max\{2, \log(\rho^{-1})\}}}$

 $= \mathbb{E}[\log|x| \log|x|^T]$

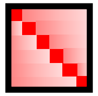
 $= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

- Apply λ_{min} to Hermite formula:

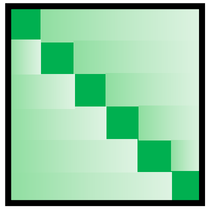
λ_{min}  $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$  $^{(2l)}$

- Apply Gershgorin Circle Theorem:

λ_{min}  $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

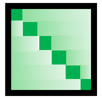
(for large enough l)

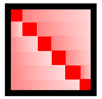
The Full, Ugly Bound

λ_{min}  $\geq$

$\sum_{l=1}^{\frac{\log(p-1)}{2\log(\rho^{-1})}} \frac{\lambda_{min}}{5l^{3/2}}$  $^{(2l)}$ $+ \frac{2}{5}\sqrt{\frac{2\log\rho^{-1}}{\log\frac{p-1}{\rho} + \max\{2, \log(\rho^{-1})\}}}$

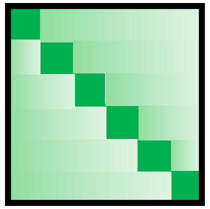
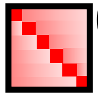
(first term, under the brace:) ≥ 0

 $= \mathbb{E}[\log|x| \log|x|^T]$

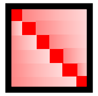
 $= \mathbb{E}[xx^T]$

Population Minimum Eigenvalue

- Apply λ_{min} to Hermite formula:

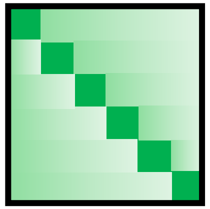
λ_{min}  $\geq \sum_{l=1}^{\infty} c_{2l}^2 \lambda_{min}$  $^{(2l)}$

- Apply Gershgorin Circle Theorem:

λ_{min}  $^{(2l)} \geq 1 - (p-1)\rho^{2l}$

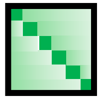
(for large enough l)

The Full, Ugly Bound

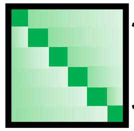
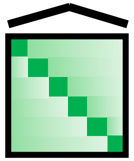
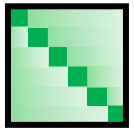
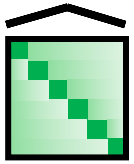
λ_{min}  $\geq$

$\sum_{l=1}^{\frac{\log(p-1)}{2\log(\rho^{-1})}} \frac{\lambda_{min}}{5l^{3/2}}$  $^{(2l)} + \frac{2}{5}\sqrt{\frac{2\log\rho^{-1}}{\log\frac{p-1}{\rho} + \max\{2, \log(\rho^{-1})\}}}$

> 0

 $= \mathbb{E}[\log|x| \log|x|^T]$

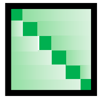
Concentration of Restricted Eigenvalue

- $|\lambda_{RE(k)}($  $)$ - $\lambda_{RE(k)}($  $)| < k \cdot ||$  $-$  $||_\infty$
  - Follows from Holder’s inequality and the Restricted Cone condition

- Log-transformed variables are **sub-exponential**

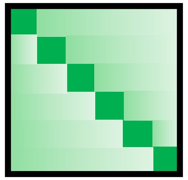
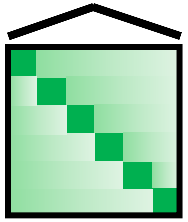
- $\max_{j,k \in [p]} \frac{1}{m} \sum_{i=1}^{m} Var\left(\log\left|x_j^{(i)}\right| \log\left|x_k^{(i)}\right|\right) \leq C$

- Elementwise ℓ_∞ error concentrates
  - [Kuchibhotla & Chakrabortty ‘18]

 $= \mathbb{E}[\log|x| \log|x|^T]$

Concentration of Elementwise ℓ_∞ error

With probability at least $1 - 3e^{-t}$,

$||$  $-$  $||_\infty$

$$\leq O\left(\sqrt{\frac{t + \log p}{m}} + \frac{(\log m)^2 (t + \log p)^2}{m}\right)$$

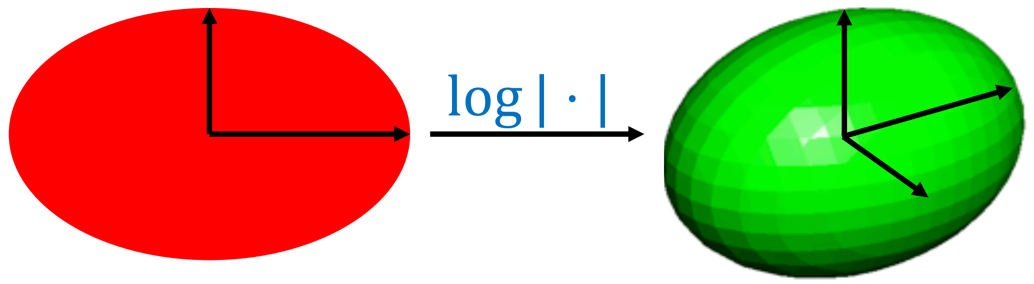
[Kuchibhotla & Chakrabortty ‘18]

4. Conclusion

Recap

- Attribute-efficient algorithm for **monomials**
  - Prior (nonlinear) work: **uncorrelated** features
  - This work: allow highly **correlated** features
    - Works beyond multilinear monomials

- Blessing of nonlinearity



Future Work

- Rotations of product distributions
- Additive noise
- Sparse polynomials with correlated features

Thanks! Questions?